Causal AI-AugETM: Integrating Structural Causal Models with Exposure–Toxicity Modeling for Robust Drug Safety Assessment Under Confounding Bias

Larry Diaz
Department of Computer Science, University of Houston, Houston, TX, USA.
larry.diaz209@uh.edu

Samuel Parker
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
samuelparker830@binghamton.edu

George Baker
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
george.work@unh.edu

Hugo Thornton
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
thornton1995@ku.edu

Abstract

The assessment of drug safety in early-phase clinical trials and post-market surveillance is persistently challenged by confounding bias arising from non-randomized treatment assignment, time-varying exposures, and unmeasured covariates. Traditional exposure-toxicity joint models, while effective in capturing dose-response relationships, lack the structural causal reasoning necessary to distinguish genuine toxicity signals from spurious associations. This paper introduces Causal AI-AugETM, a comprehensive framework that integrates structural causal models with an AI-augmented exposure-toxicity joint modeling paradigm. The framework systematically embeds causal graphs, intervention calculus, and counterfactual reasoning into the exposure-toxicity modeling pipeline, enabling robust estimation of causal effects under complex confounding scenarios. Emphasis is placed on the system-level architecture that governs data ingestion, causal discovery, model inference, and feedback loops for continuous updating. The design explicitly addresses structural trade-offs between model flexibility and interpretability, computational feasibility and statistical precision, as well as fairness across demographic subgroups that may be differentially affected by confounding by indication. Furthermore, the paper discusses governance and policy implications, including regulatory validation standards, transparency requirements, and accountability mechanisms for AI-driven safety assessments. Deployment considerations such as scalability via cloud-native and federated infrastructures, sustainability of model retraining cycles, and integration with existing pharmacovigilance systems are examined. Cross-domain comparisons with causal inference applications in climate science and econometrics provide additional insight into best practices. The Causal AI-AugETM framework represents a principled step toward trustworthy, causally informed drug safety assessment that can adapt to evolving real-world data environments while maintaining scientific rigor and societal trust.

Keywords

causal inference, structural causal models, exposure-toxicity modeling, drug safety, confounding bias, AI-augmented clinical trials, robust assessment, socio-technical infrastructure, fairness, governance.

1. Introduction

Drug safety assessment remains one of the most critical and methodologically challenging phases of pharmaceutical development. Early-phase clinical trials, while tightly controlled, are often limited by small sample sizes and short follow-up periods, making it difficult to detect rare or delayed toxicities. Post-market observational studies, in contrast, offer larger and more diverse populations but are notoriously susceptible to confounding bias because treatment assignment is not randomized and is often correlated with patient health status, comorbidities, and concomitant medications [1]. Traditional exposure-toxicity joint models, which simultaneously model the longitudinal trajectories of drug exposure and biomarker-based toxicity, have improved the ability to characterize dose-response surfaces [2]. However, these models lack formal causal machinery and thus cannot reliably separate the effect of the drug from the effect of confounding variables that drive both exposure and toxicity.

Recent advances in artificial intelligence, particularly in deep learning and high-dimensional pattern recognition, have been leveraged to augment exposure-toxicity modeling by capturing complex nonlinear interactions [3,4]. Yet these AI-augmented approaches, if applied without a causal grounding, may amplify biases present in the training data or learn spurious correlations that are not transportable across populations or dosing regimens [5]. To bridge this gap, we propose Causal AI-AugETM, a framework that integrates structural causal models (SCMs) with AI-augmented exposure-toxicity joint modeling to provide robust and interpretable causal effect estimates under confounding. This integration is not merely an algorithmic addition; it requires a fundamental rethinking of the system architecture, data governance, and deployment strategies that underlie modern pharmacovigilance. The present paper adopts a system-level perspective, examining the structural trade-offs inherent in this design—such as the tension between model complexity and causal identifiability, the computational burden of counterfactual simulation, and the need to ensure fairness across diverse patient subgroups that may experience differential confounding [6]. We also explore the policy and regulatory implications of deploying causally informed AI systems for drug safety, including standards for validation, transparency, and accountability.

2. Background and Motivation

The concept of exposure-toxicity joint modeling originates from the need to account for the dynamic interplay between drug exposure (e.g., dose, concentration, cumulative exposure) and toxicity endpoints, which are often time-to-event or longitudinal biomarkers [7]. Classical joint models typically employ shared random effects to link the exposure process and the toxicity outcome, but they assume a purely associational relationship that can be severely biased when confounders affect both the exposure trajectory and the risk of toxicity [8]. For example, in oncology, patients with more aggressive disease may receive higher doses while also having a higher baseline risk of adverse events, creating a non-causal correlation that naive joint models would misinterpret as a direct toxic effect.

Structural causal models, as formalized by Pearl [9] and extended by others [10], provide a rigorous language for encoding causal assumptions via directed acyclic graphs, intervention calculus, and counterfactual logic. In the context of drug safety, SCMs allow researchers to

specify which variables are common causes of exposure and toxicity, and then compute the causal effect of a particular dosing regimen through hypothetical interventions, even when randomization is infeasible [11]. The integration of SCMs with exposure-toxicity modeling has been explored in limited settings, but most implementations rely on parametric assumptions and cannot easily handle high-dimensional confounders such as electronic health records, genomic data, or real-world evidence streams [12].

AI-augmented methods, including deep learning and ensemble models, have been increasingly applied to exposure-toxicity modeling to capture nonlinear dose-response surfaces and interactions among many covariates [13,14]. One notable framework is the AI-AugETM approach, which uses neural networks to jointly model exposure and toxicity trajectories while employing regularization and adversarial training to mitigate certain forms of bias [15,16]. However, the AI-AugETM framework, as originally conceived, does not explicitly embed a structural causal model; its bias reduction relies on data-driven corrections that may fail when confounding patterns shift across populations or time [17]. This limitation motivates the need for a principled integration: Causal AI-AugETM builds upon the infrastructure of AI-augmented joint modeling but grounds it in a causal graph that defines the relationships among exposures, confounders, and outcomes. This grounding enables the use of formal identification strategies, such as front-door and back-door adjustment, instrumental variables, and g-estimation, which can be executed with the computational power of deep learning while maintaining causal interpretability [18].

3. Methodological Framework of Causal AI-AugETM

The Causal AI-AugETM framework is organized around three interconnected layers: causal structure specification, exposure-toxicity joint modeling with causal constraints, and counterfactual simulation for inference and validation. At the foundation, a causal directed acyclic graph is constructed either from domain knowledge, literature, or causal discovery algorithms that leverage the available data and prior constraints [19]. This graph encodes the assumed relationships between measured and unmeasured confounders, the exposure variable (which may be time-varying), and the toxicity outcome. In many drug safety scenarios, confounders include disease severity, renal function, genetic markers, and concurrent medications. The graph is then used to derive identification formulas for the causal effect of a hypothetical dose regimen on toxicity risk, typically through the do-operator or counterfactual probabilities.

The second layer consists of an AI-augmented joint model that learns the conditional distributions specified by the causal graph. Instead of fitting separate models for exposure and toxicity, the framework employs a shared deep representation that respects the conditional independence assumptions implied by the graph. For instance, if the graph implies that exposure and toxicity are conditionally independent given confounders and past exposure history, the joint model is regularized to enforce this constraint, reducing the risk of learning spurious associations [20]. The neural network architecture is designed to accommodate longitudinal data with missingness, irregular sampling, and high-dimensional covariates. Importantly, the model outputs are not merely predictive but are structured to support causal queries: for a given patient covariate profile, the model estimates the counterfactual toxicity risk under a counterfactual dose trajectory, obtained by intervening on the exposure node while keeping confounders fixed.

The third layer implements counterfactual simulation at scale. Using the estimated joint model as a generative engine, the framework draws samples from the conditional distributions

under different intervention scenarios. These samples are used to compute average causal effects (e.g., the population-level risk difference between two dosing regimens) and conditional average effects for subgroups defined by age, sex, comorbidities, or genetic markers. Uncertainty quantification is performed via bootstrapping or variational inference, producing credible intervals that account for both model and sampling uncertainty [21]. The entire pipeline is designed to be modular, allowing the causal graph to be updated as new evidence emerges, and the AI model to be retrained with new data without rebuilding the entire architecture.

4. Structural Design and System Architecture

From an architectural viewpoint, Causal AI-AugETM is deployed as a layered system that spans data ingestion, causal graph management, model training, inference, and continuous monitoring. The data ingestion layer must handle heterogeneous sources: electronic health records, pharmacokinetic data, adverse event reports, laboratory results, and possibly genomic or wearable sensor data. Each source has its own missing data patterns, measurement error structures, and temporal granularities. The system must implement harmonization and imputation strategies that are informed by the causal graph to avoid introducing spurious dependencies [22]. For example, missingness not at random can be treated as a node in the graph, enabling a principled adjustment.

The causal graph management layer is a critical governance component. It stores the current best knowledge of causal relationships, along with uncertainty bounds and evidence provenance. This graph is not static; it must be revisited as new domain knowledge emerges or as causal discovery algorithms suggest modifications. The system supports version control, allowing users to roll back or compare alternative graph structures. Automated causal discovery algorithms, such as the PC algorithm or fast causal inference, are integrated but their outputs are flagged for human review [10]. This hybrid human-AI approach mitigates the risk of over-reliance on automated graph learning, which can be unreliable in high-dimensional settings with unmeasured confounders.

The model training and inference layer orchestrates the AI-augmented joint model. Because the causal structure imposes constraints, training involves multiple loss functions: one for fitting the observed data distribution, and additional penalty terms that enforce conditional independence relationships implied by the graph. This constrained optimization can be computationally intensive, especially when the graph includes time-varying dependencies or latent confounders. To manage computational cost, the architecture supports distributed training across GPU clusters, and uses mini-batch approximations with careful tuning to avoid convergence to biased solutions. Inference for counterfactual queries is performed via Monte Carlo sampling from the trained generative model, which can be parallelized effectively.

Feedback loops are incorporated to enable adaptive learning. As new data accumulate from ongoing trials or post-market surveillance, the system updates the model parameters and, if warranted, the causal graph. However, continual updating presents a trade-off between responsiveness and stability: too frequent updates may induce volatility in causal estimates, while too infrequent updates may allow model drift. The architecture implements a monitoring dashboard that tracks key performance indicators such as the balance of confounders across exposure groups, the degree of violation of graph-implied conditional independencies, and the stability of effect estimates over time [23]. Thresholds trigger human review before automatic retraining is allowed, ensuring that governance remains in the loop.

5. Trade-offs in Confounding Mitigation and Model Robustness

Any system designed to mitigate confounding bias faces fundamental trade-offs. One central tension is between the complexity of the causal model and the identifiability of causal effects. A more detailed causal graph that includes many potential confounders, mediators, and latent variables may better approximate reality, but it also increases the number of assumptions that must be justified and the data requirements for estimation [24]. Under-specified graphs, by contrast, may omit important confounders and yield biased estimates. In Causal AI-AugETM, this trade-off is managed by using a hierarchy of models: a baseline graph with core confounders derived from clinical guidelines is always fitted, while extensions to include additional covariates or latent confounders are treated as sensitivity analyses. Uncertainty intervals are widened to reflect the dependence of estimates on untestable assumptions.

Another trade-off concerns the robustness of the AI-augmented joint model to misspecification of the causal graph. If the graph is incorrect—for example, if an unmeasured confounder is omitted—the model may still produce seemingly stable estimates that are actually biased. To address this, the framework integrates multiple causal estimators (e.g., inverse probability weighting, doubly robust estimation, g-computation) and compares their results; divergence between estimators signals possible model misspecification [25]. Additionally, a range of sensitivity analyses are automated, such as the introduction of a hypothetical unmeasured confounder with variable strength, allowing users to assess how severe a confounder must be to reverse the conclusion.

Fairness is a cross-cutting concern. Confounding often systematically differs across demographic groups, leading to biased toxicity estimates for minorities or underrepresented populations. For instance, if a certain ethnic group has different baseline renal function on average, and renal function confounds both exposure and toxicity, then ignoring this interaction can lead to under- or over-estimation of risk. The system architecture permits subgroup-specific causal graphs and models, but this introduces a trade-off between group-level precision and sample size. A stratified approach may yield unstable estimates for small subgroups, while a pooled approach may impose uniform assumptions that are inaccurate. The framework employs Bayesian hierarchical modeling to share information across subgroups while allowing for heterogeneity, and it reports both pooled and subgroup-specific effect estimates with clear uncertainty statements [6]. Policy recommendations explicitly discuss the fairness implications of using one model over another, recognizing that technical choices have real-world consequences for patient safety and equity.

6. Governance, Fairness, and Policy Implications

The deployment of Causal AI-AugETM within regulatory and clinical decision-making contexts raises significant governance challenges. Current regulatory frameworks for drug safety, such as those from the FDA and EMA, have begun to acknowledge the role of causal inference but have not yet established clear standards for AI-assisted causal effect estimation [26]. The framework is designed to be compliant with existing good pharmacovigilance practices by maintaining a human-in-the-loop for critical decisions: all final safety assessments based on causal estimates must be reviewed by clinical experts, and the system records the full chain of assumptions, graph versions, and model outputs to support auditability.

Transparency is essential for trust. The system provides a causal explanation for each estimated effect, visualizing the graph and highlighting the paths through which confounding

was adjusted. This feature is particularly important for communicating with clinicians and regulators who may be skeptical of black-box AI models. Furthermore, the framework includes a fairness audit module that automatically tests for differential impact across protected groups. If, for example, a dosing recommendation leads to higher predicted toxicity in one demographic group compared to another under the same causal estimate, the system flags this and prompts a deeper investigation into whether confounding or measurement bias is responsible [27]. Policy recommendations include requiring that any AI-augmented drug safety system be validated on diverse populations and that its graph assumptions be publicly registered before deployment.

From a broader socio-technical perspective, the governance of Causal AI-AugETM must also address data privacy and ownership. The model relies on potentially sensitive patient data, and the counterfactual simulation engine can generate synthetic patient-level data that, if misused, could violate privacy. The architecture therefore incorporates differential privacy mechanisms during training and limits the granularity of output summaries. Moreover, federated learning paradigms are supported, allowing multiple hospitals to jointly train the model without sharing raw patient records [28]. This federated deployment, however, introduces additional complexity in causal graph consistency across sites, as the causal structure may differ due to local clinical practices or population characteristics. The governance structure must establish protocols for reconciling graph differences and for ensuring that site-specific effects are not inadvertently masked by pooled estimates.

7. Deployment, Scalability, and Sustainability

Deploying Causal AI-AugETM at scale requires a robust, cloud-native infrastructure capable of handling the computational demands of causal inference with deep generative models. The system is designed as a set of microservices, each responsible for a distinct function: data preprocessing and imputation, causal graph management, joint model training, counterfactual inference, and monitoring. Containerization and orchestration via Kubernetes enable elastic scaling; for example, during a batch of counterfactual simulations for a large trial, additional worker pods can be spun up to handle the workload, then released. This architecture also facilitates continuous integration and deployment, allowing updates to the causal graph or the neural network architecture to be rolled out with minimal downtime.

Scalability is not only about computational resources but also about maintaining model quality as data volume grows. The framework implements an incremental learning strategy, where the joint model is updated using new data without full retraining, provided that the causal graph remains unchanged. When the graph is revised, a full retraining is triggered, but the system can leverage transfer learning from the previous model to accelerate convergence. Sustainability concerns include the energy consumption of repeated training cycles, especially as deep generative models become larger. To mitigate this, the system employs early stopping, model pruning, and quantization, and it schedules resource-intensive jobs during off-peak hours using green energy when available [29]. Furthermore, the framework is designed to be modular so that lighter-weight causal models (e.g., linear structural equation models) can be substituted for the deep learning component in low-resource settings, albeit with a potential loss of expressiveness. This flexibility supports deployment in low- and middle-income countries where computational infrastructure may be limited, promoting global equity in drug safety assessment.

Integration with existing pharmacovigilance systems is critical for adoption. Causal AI-AugETM exposes standardized APIs (e.g., RESTful interfaces) that can be consumed by

electronic health record systems, clinical trial management platforms, and adverse event databases. The output is formatted as structured reports that include the estimated causal effects, confidence intervals, sensitivity analyses, and a summary of assumptions. These reports can be directly ingested into regulatory submission pipelines. The system also supports automated alerts: if a newly observed toxicity pattern is causally attributable to a particular dose regimen beyond a pre-specified threshold, an alert is generated for pharmacovigilance teams.

8. Case Studies and Cross-Domain Comparisons

To illustrate the practical application of Causal AI-AugETM, consider a hypothetical but realistic scenario in oncology: evaluating the cardiac toxicity of a novel targeted therapy across a multi-site trial where dose adjustments are made based on renal function and previous adverse events. Traditional joint models that ignore the feedback between dose and toxicity would overestimate the toxic effect because patients with declining renal function both receive lower doses and are at higher risk of cardiac toxicity. Using the proposed framework, the causal graph explicitly models renal function as a time-varying confounder that affects both dose and toxicity, and also includes previous toxicity as a mediator of future dose decisions. The AI-augmented joint model then learns the conditional distributions, and counterfactual simulation estimates what the toxicity risk would be under a strategy that does not adjust dose based on renal function, revealing the true drug effect. This case demonstrates how the framework handles time-varying confounding, a notoriously difficult problem in causal inference.

Cross-domain comparisons provide additional insight. In climate science, structural causal models have been used to attribute extreme weather events to anthropogenic forcing, with careful handling of spatial and temporal confounding [30]. The parallels with drug safety are striking: both fields deal with non-experimental data, complex feedback loops, and high-stakes policy decisions. Lessons learned from climate attribution—such as the importance of ensemble methods and the need to communicate uncertainty clearly—directly inform the design of Causal AI-AugETM. Likewise, econometric methods for causal inference, particularly the use of instrumental variables to address selection bias, offer alternative identification strategies that could be incorporated into the framework for scenarios where a valid instrument exists (e.g., genetic variants as instrumental variables in pharmacogenomics). The modular architecture allows such strategies to be plugged in as needed.

9. Future Directions

The current architecture of Causal AI-AugETM opens several avenues for future research and development. One promising direction is the integration of reinforcement learning for dynamic dosing optimization. Instead of only estimating causal effects of fixed dose trajectories, the system could learn an optimal dosing policy that balances efficacy and toxicity, using counterfactual value estimation to avoid confounding bias during policy evaluation [31]. This would require extending the causal graph to include a decision node representing the current dosing rule and incorporating off-policy correction techniques.

Another area is the automated discovery of causal graphs from large-scale observational data, especially when domain knowledge is incomplete. Modern causal discovery algorithms that incorporate domain adaptation and handling of latent confounders are increasingly reliable, but their outputs should be combined with human expertise [19]. The framework could implement an active learning protocol, where the system queries clinicians about specific

edges in the graph to resolve ambiguity. Additionally, the use of large language models to extract causal relationships from medical literature could be employed to initialize the graph, with subsequent refinement based on data.

Finally, the governance and regulatory aspects require ongoing collaboration with agencies such as the FDA, EMA, and health technology assessment bodies. Establishing a standardized reporting template for causal effect estimates from AI-augmented models would facilitate review and approval. The framework's emphasis on transparency and sensitivity analysis positions it well to contribute to emerging regulatory frameworks for AI in healthcare.

10. Conclusion

Causal AI-AugETM represents a comprehensive system-level integration of structural causal models with AI-augmented exposure-toxicity joint modeling, addressing the persistent challenge of confounding bias in drug safety assessment. By embedding causal graphs, intervention calculus, and counterfactual simulation into a scalable, cloud-native architecture, the framework provides robust and interpretable causal effect estimates that can inform dose optimization and risk management. The design explicitly grapples with structural trade-offs between complexity and identifiability, fairness and precision, and computational cost and statistical accuracy. Governance mechanisms ensure transparency, auditability, and human oversight, while deployment strategies support federated learning and resource-adaptive configurations. Cross-domain comparisons with climate attribution and econometrics enrich the methodological toolkit, and future directions point toward dynamic policy learning and automated graph discovery. As pharmaceutical development increasingly relies on real-world evidence and AI, principled causal frameworks such as Causal AI-AugETM will be essential to maintaining scientific rigor, regulatory trust, and equitable patient outcomes.

References

1. Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if. Chapman & Hall/CRC.
2. Tsiatis, A. A., & Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica, 14(3), 809–834.
3. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.
4. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.
5. D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alaa, A., Beutel, A., ... & Zhang, L. (2020). Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(1), 1–61.
6. Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in Neural Information Processing Systems (pp. 4066–4076).
7. Rizopoulos, D. (2012). Joint models for longitudinal and time-to-event data with applications in R. CRC Press.
8. Daniel, R. M., Cousens, S. N., De Stavola, B. L., Kenward, M. G., & Sterne, J. A. C. (2013). Methods for dealing with time-dependent confounding. Statistics in Medicine, 32(9), 1584–1618.
9. Pearl, J. (2009). Causality (2nd ed.). Cambridge University Press.

10. Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, prediction, and search (2nd ed.). MIT Press.

11. Robins, J. M., & Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In Advances in longitudinal data analysis (pp. 553–599). CRC Press.

12. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences, 113(27), 7353–7360.

13. Li, Y., Wang, J., Ye, J., & Reddy, C. K. (2017). A multi-task learning formulation for survival analysis. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1715–1724).

14. Ranganath, R., Perotte, A., Elhadad, N., & Blei, D. M. (2016). Deep survival analysis. In Machine Learning for Healthcare Conference (pp. 101–114).

15. Zhang, Z., Chen, L., & Xu, Y. (2023). Neural joint models for exposure-toxicity analysis with high-dimensional covariates. Journal of Biomedical Informatics, 140, 104321.

16. Wang, Y. (2025, August). AI-AugETM: An AI-augmented exposure–toxicity joint modeling framework for personalized dose optimization in early-phase clinical trials. In 2025 19th International Conference on Complex Medical Engineering (CME) (pp. 182-186). IEEE.

17. Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In International Conference on Machine Learning (pp. 3020–3029).

18. van der Laan, M. J., & Rose, S. (2011). Targeted learning: Causal inference for observational and experimental data. Springer.

19. Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10, 524.

20. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. arXiv preprint arXiv:1907.02893.

21. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC Press.

22. Molenberghs, G., & Kenward, M. G. (2007). Missing data in clinical studies. Wiley.

23. Saria, S., & Subbaswamy, A. (2019). Tutorial: Safe and reliable machine learning. arXiv preprint arXiv:1904.07204.

24. Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling, 7(9–12), 1393–1512.

25. Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962–973.

26. U.S. Food and Drug Administration. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. FDA.

27. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226).

28. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

29. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645–3650).

30. Hannart, A., Pearl, J., Otto, F. E. L., Naveau, P., & Ghil, M. (2016). Causal counterfactual theory for the attribution of weather and climate events. Journal of Climate, 29(8), 3001–3024.